Statistical Process Control for Flight Delays of American Airlines Incorporation

Student ID:40028977 - saeed keshavarz- INSE 62220

***Abstract*— Abstract—Airport congestion and flight delays is essence of transportation in North America. Flight delays are one of the most common customer complaints reported by airline passengers. The cost of flight delays is not only restricted to airlines, but also a delayed flight can be costly to passengers, thus having a strong impact on quality of airline. To improve quality of an airline, flight delays should be controlled. In this project, statistical process control, using Principal Component Analysis with Logistic regression, is implemented to classify flight delays by analyzing principal components and latent factors.**

Index Terms—Flight Delays, Statistical Process Control, Prin- cipal Component Analysis, Logistic regression

I. INTRODUCTION

A. *Flight Delays and Airline Quality*

FLIGHT DELAY is one of the key factors which define Quality of an Airline. In the United States, the Federal Aviation Administration estimates that flight delays cost airlines 22 billion annually [12]. According to FAA, a flight is considered delayed when it arrived 15 or more minutes than the schedule. The cost of flight delays is also related to passengers, as due to a delayed flight, they may not be able to meet their schedules. Also, a passenger who is delayed on a multi-plane trip could miss a connecting flight.

A study conducted by researchers of University of California, Berkeley found that the cost of domestic flight delays puts a 32.9 billion dent into the U.S. economy, and about half that cost is borne by airline passengers [10].

In this project, flight delays of American Airline Incorporations, at Chicago O'Hare International Airport (ORD) Airport, Chicago, USA, which s the busiest Airpot in the U.S are analyzed for years 2016 and 2019. Total 37 months' data, i.e. from Sep 2016 to Sept 2019, is observed with 5 delaying factors (variables). The delaying factors are described as [9];

Air Carrier: The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).

Extreme Weather: Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.

National Aviation System (NAS): Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.

Security: Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.

Late-arriving aircraft: A previous flight with same aircraft arrived late, causing the present flight to depart late.

The Flight Delay data, shown in Table 1, is taken from Bureau of Transportation Statistics, Research and Innovative Technology Asministration, United States Department of Transportation [9].

The Fig.1 shows the sample of our elements with features.

Column1	Carrier	weather	nas	security	late aircraft
D1	57651	3342	20147	64	16106
D2	42310	192	11417	64	13771
D3	46997	640	16101	85	15236
D4	68522	2917	19234	194	24546
D5	47648	2996	14346	111	16096
D6	35061	1488	8780	26	9565
D7	60422	1858	19112	235	23249
D8	55313	1704	14677	103	18531
D9	63469	1579	14624	3	24255
D10	93694	5412	33136	73	31742
D11	111483	6934	36743	415	40654

Fig. 1. Sample rows of Flight Delays

B. *Principle Component analysis*

Principal Component Analysis (PCA) is a multivariate statistical technique used for transforming a set of observations of possibly correlated multiple variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is normally less than the number of original variables. The first principal component has the largest possible variance, i.e. it accounts for as much of the variability in the data as possible, and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components. Principal components are guaranteed to be independent only if the data set is jointly normally distributed. If a multivariate dataset is visualised as a set of coordinates in a high-dimensional data space (1 axis per variable), PCA can supply the user with a lower-dimensional picture, a "shadow" of this object when viewed from its (in some sense) most informative viewpoint. This is done by using only the first few principal components so that the dimensionality of the transformed data is reduced [1]. The objective of PCA is to reduce the dimensionality of a dataset while retaining as much information as possible to the inherent variability within the data. Standardizing the data is often preferable when the variables are in different units or when the variance of the different columns of the data is

substantial [3]. Given a data matrix X, the PCA algorithm consists of four main steps:

Step 1: Compute the centered data matrix Y = HX by subtracting off-column means.

Step 2: Compute the p p covariance matrix S of the centered data matrix Step 3: Compute the eigenvectors and eigenvalues of S using eigen-decomposition Step 4: Compute the transformed data matrix Z = YA of size n p

which contains the coordinates of the original data in the new coordinate system defined by the PCs. The rows of Z correspond to observations zi = A(xi x), while its columns correspond to PC scores.

C. Logistic regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression[1] (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled ”0” and ”1”. In the logistic model, the log-odds (the logarithm of the odds) for the value labeled ”1” is a linear combination of one or more independent variables (”predictors”); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled ”1” can vary between 0 (certainly the value ”0”) and 1 (certainly the value ”1”), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.[13]

II. PROBLEM DESCRIPTION

In this Project we want to see if we can predict the level of Delays in the future month. For this purpose first we Need decrease the dimension of our data to two element With help of PCA and in the next step use Logistic regression to see if this algorithm is right approach for this problem.

First we have standardize our data before feeding it to PCA and after running PCA we used our labeled data that has two classes(class 1 high traffic month, class 2 low traffic month) to run prediction test.

III. SOLUTION

A. Box Plot for Delaying Factors

Boxplots of delaying factors are shown in Fig. 2. It can be observed from the figure that there are few outliers for security feature. Also the impact of each variable (delaying factor) is not highly different. NAS and Weather Delay, the thirdand second variable similarly have the highest impact whereas, Security Delay, the fourth variable, has the lowest impact. Also, the means of variables are the similar but not the variance. Late Aircraft Delay has the largest mean and Weather largest variance whereas, Security Delay has the lowest.

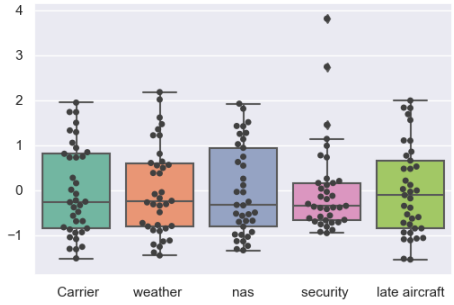

Fig. 2. Box Plot of Flight Delays

B. Correlation Matrix of Delaying Factors

Correlation Matrix of Delaying Factors Correlation Matrix of delaying factors is shown in Fig. 3. It can be observed from the figure that most of the variables are highly correlated with each other. Carrier Delay is correlated with all other variables. NAS delay has almost same and strong correlation with all other variables except Security Delay. Security Delay has the least correlation with all other variables specially NAS delay.

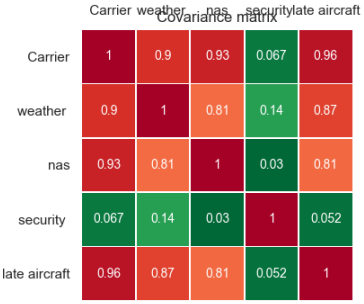

Fig. 3. Covariance Matrix of Delaying Factors

C. Scatter plot Matrix of Delaying Factors

When interpreting correlations it is important to visualize the bivariate relationships between all pairs of variables. From the scatter plot matrix of delaying factors, shown in Fig. 4, it can be observed that bivariate relationship exists between delaying factors except Security Delay.

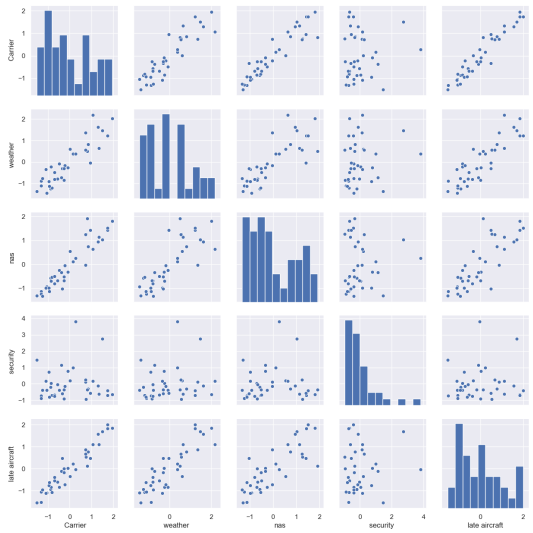

Fig. 4. Correlation Matrix of Delaying Factors

IV. PCA IMPLEMENTATION

Next step in PCA analysis is to implement the eigen value decomposition of the covariance matrix Using the columns of the eigenvector matrix;

$$A = \begin{pmatrix} -0.4880 & 0.0063 & -0.6844 & 0.0887 & 0.5344 \\ -0.4623 & -0.3259 & 0.6711 & 0.2695 & 0.3964 \\ -0.4894 & -0.1627 & 0.0546 & -0.8210 & -0.2388 \\ -0.2382 & 0.9312 & 0.2505 & -0.0554 & 0.1015 \\ -0.5018 & 0.0108 & -0.1248 & 0.4924 & -0.7000 \end{pmatrix}$$

The first PC is given by:

Z1 = 0.4880X1 0.4623X2 0.4894X3 0.2382X4 0.5018X5

Z1 = 0.4880CD0.4623WD0.4894ND0.2382SD0.5018LD

We can see that the first PC contains negative values for all delaying factors. So, they all are correlated. The second PC is given by

Z2 = 0.3259X2 0.1627X3 + 0.9312X4

Z2 =0.3259WD0.1627ND+0.9312SD

We can see that the second PC is mainly a contrast between Security Delay and other delaying factors like Weather Delay and NAS Delay. This is evidenced by the positive coefficient for security delay and the negative coefficients for weather delay and NAS delay.

Scatter plot of PC2 Score Vs. PC1 Score is shown in Fig. 4. It can be observed that the observation of Aug 2012 is exceptionally different from others. From Table 1, it can be verified that in this month, 36 delays due to Security occured which are exceptional.

A. Principal Component score

Scatter plot of PC2 Score Vs. PC1 Score is shown in Fig. 5. It can be observed that the observation of Month 11 and Month 25th in data set is exceptionally different from others and has high delays mainly related to security delay occurrence.

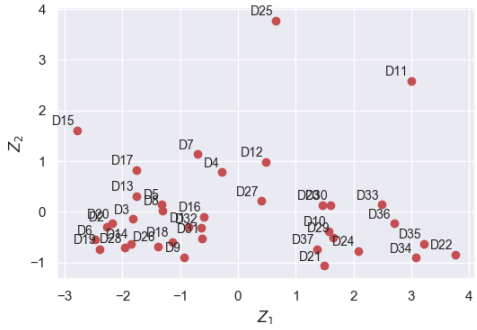

Fig. 5. Scatter plot of PC2 Score Vs. PC1 Score

B. Principal component Coefficient

The scatter plot of PC2 coefficients vs. PC1 coefficients is shown in Fig. 6. This plot helps understand which variables have a similar involvement within PCs. As it can be observed from figure, that the Security Delay is plotted far away from other delaying factors and has the maximum magnitude, whereas all other delaying factors are plotted close to each other. It shows the strong relation of PC1 and PC2 in terms of Security Delay and weak relation in terms of Weather Delay.

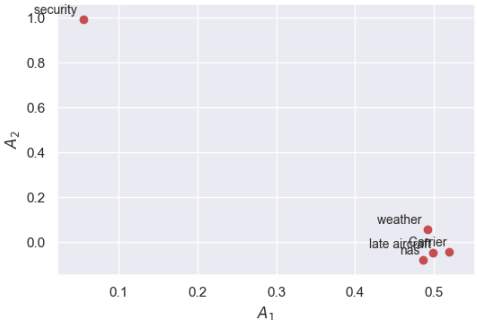

Fig. 6. Scatterplot of PC2 Coefficient Vs. PC1 Coefficient

The scatterplot of PC3 coefficients vs. PC2 coefficients is shown in Fig. 7. It can be observed that PC 2 and PC3 have strong relation in terms of Weather Delay and weak relation in terms of NAS Delay.

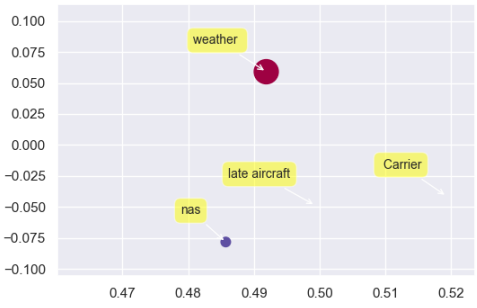

Fig. 7. Scatter plot of PC2 Coefficient Vs. PC3 Coefficient

C. Explained Variance and Lowest-Dimensional Space

The percentage of variance accounted for by the jth PC is called explained variance, and it is given by;

$$l_j = \frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j} * 100\%, \quad j = 1, ..., p$$

From below value of explained variances in Fig. 8, it can be ob- served that PC1 and PC2 combined account for roughly 90percent of the variance in the data. Therefore, based upon the explained variance by both PC1 and PC2 and also from the Scree and Pareto plots, it can be deduced that the lowest-dimensional space to represent the Flight Delay data corresponds to d = 2.

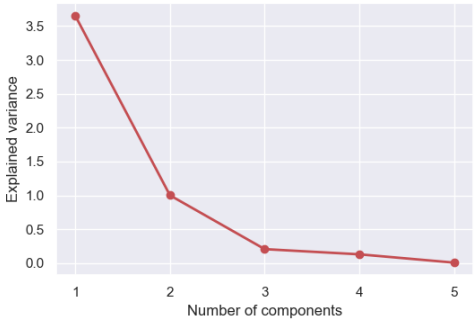

Fig. 8. scree cahrt of Delaying Factors

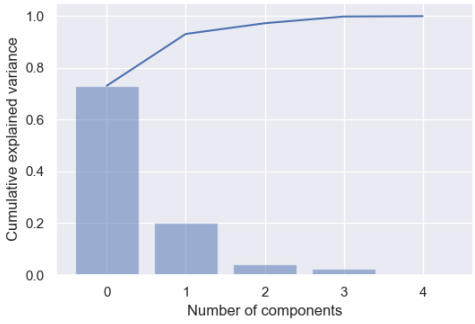

Fig. 9. Pareto cahrt of Delaying Factors

D. Biplot of Principal Components

The biplot is a graphical tool that allows information on both observations and variables of a data matrix to be dis- played graphically. Observations are displayed as points while variables are displayed as vectors. The biplot helps visualize both the principal component coefficients for each variable and the principal component scores for each observation in a single plot. Each of the p variables is represented in the biplot by a vector, and the direction and length of the vector indicates how each variable contributes to the two principal components in the biplot. The axes in the biplot represent the principal components, and the observed variables are represented as vectors.

Through biplot, we can visualize the magnitude and sign of each variable's contribution to the first two or three principal components, and how each observation is represented in terms of those components. Each of the n observations is represented in the biplot by a point, and their locations indicate the score of each observation for the two principal components in the plot.

Fig. 10 shows the 2D Biplot for Flight Delay data. From the 2D Biplot, it can be observed that some observations, points near the left edge of this plot, have the lowest scores for the first principal component. Also, vector of Security Delay has the highest magnitude and positive direction which verifies the strong relation of PC1 and PC2 in terms of Security Delay, which has also been observed in Fig. 6. The direction of vector of NAS adn Delay verifies the weakness of relation of PC1 and PC2 in terms of NAS Delay, which has also been observed in Fig. 6.

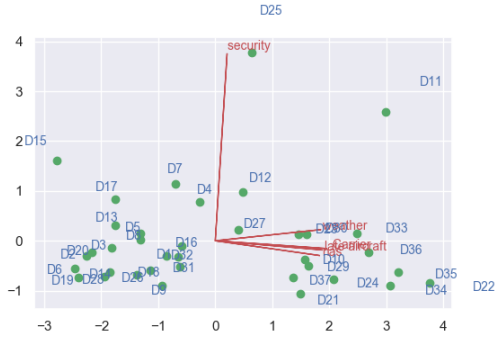

Fig. 10. 2D Biplot

In PCA correlation matrix Fig.11. we can s the strongest relations that exist in first two component and adn the weakest and strongest in weather in 3th component and 1th

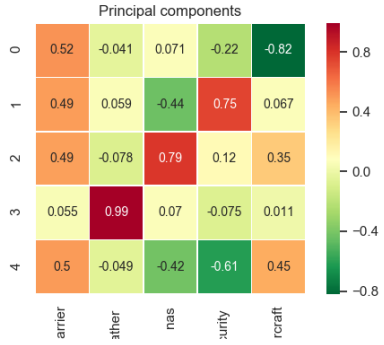

Fig. 11. Correlation Matrix of PCA

V. Logistic Regression

After Finding the best 2 Component we ran logistic regression on our data, the label that we used is hand made, since the raw data had not label; we split ted our data with train test split method adn we used 80-20 percent approach, in the below figures we plotted the confusion matrix.

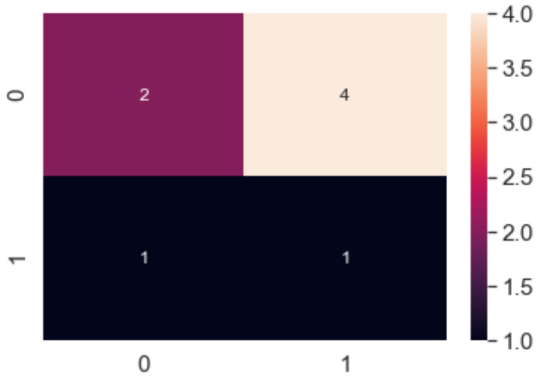

Fig. 12. Confusion Matrix

Our approach with Logistic regression had nearly 40 percent accuracy , which shows based on the current setting is not the Best algorithm for prediction, in that order we plotted the shape of trained and test data to see the problem, adn based on the below figures I believe the SVM is a better approach to predict adn also we can use more month to feed our data set.

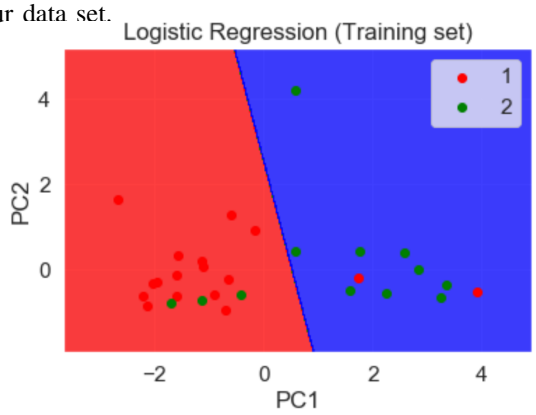

Fig. 13. train scatter plot

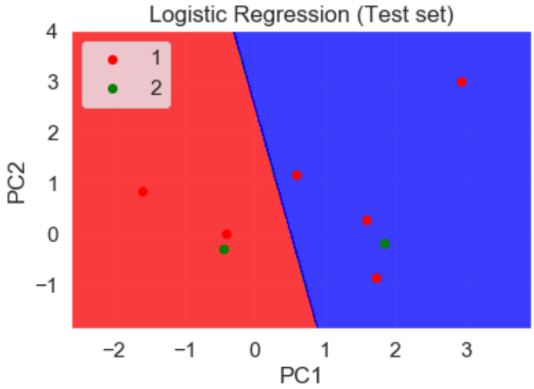

Fig. 14. test scatter plot

VI. CONCLUSION

In this project, through the implementation of Advanced Statistical Techniques, Principal Component Analysis and Machine learning algorithm Logistic regression, on Flight Delay data, it can be concluded that American Airlines Incorporation can easily control flight delays by analyzing either the first two principal components , rather than analyzing the whole data which can be a cumbersome task. As, it is proved that the first two Principal Components contains almost 90 percent of variability and also Machine learning algorithm can predict with enough data more than what I use to predict future flight delays but Logistic regression based these approach is not the most appropriate algorithm based on the shape of the data that we have SVM is more likely to be accurate for flight prediction delays, so it is better to focus on the principal components and ML algorithm rather than analyzing the data of several years at multiple airports.

REFERENCES

[1] Stefatos, G. Ben Hamza, A. (2010). Dynamic inde- pendent component analysis approach for fault detection and diagnosis. Expert Systems with Applications, 37, 8606–8617
[2] Principal Copmonent Analysis, Wikipedia (Accessed on Dec 01, 2012).
[3] Brown, J. D. (2009). Principal components analysis and exploratory factor analysis – Definitions, differences and choices.
[4] Factor Analysis, Wikipedia (Dec 02, 2012), Internet: http://en.wikipedia.org/wiki/Factoranalysis
[5] Williams, B., Onsman, A. and Brown, T.. Exploratory factor analysis: A five-step guide for novices.
[6] Williams, B., Onsman, A. and Brown, T.. Exploratory factor analysis: A five-step guide for novices.
[7] How to Interpret Factor Analysis Results, eHow (Accessed on Dec 02, 2012).
[8] Field, A. (2000). Discovering Statistics using SPSS for Windows. Sage Publications.
[9] Data Source; Bureau of Transportation Statistics, Research and Innovative Technology Administration, United States Department of Transportation. Internet:
[10] Flight Delays Cost, News Center, University of Cal-ifornia at Berkeley (Accessed on Dec 03, 2012). Internet:
[11] Forbes, S. (2008). The effect of air traffic delays on air- line prices. International Journal of Industrial Organization, 26 (2008), 1218–1232
[12] Flight Delay, Wikipedia (Accessed on Dec 03, 2012).
[13] https://en.Wikipedia.org/wiki/Logistic$_{r}egression$